



# Technology Work in EUDAT DAITF Idea

Peter Wittenburg, Daan Broeder  
The Language Archive, Max Planck Institute, Netherlands  
DATA2012, Indianapolis



Date: 26<sup>th</sup> January 2012

# Outline of the talk

- ❑ what are the needs of communities wrt common services
  - ❑ which are the enabling technologies
- ❑ Understanding the Collaborative Data Infrastructure
  - ❑ how is data organized in communities
  - ❑ how are data centers organized (not today)
- ❑ Main Service Cases
  - ❑ Safe replication, Data Staging to HPC
  - ❑ Joint Metadata Domain, Simple Data Store
  - ❑ Federation Scalability (volumes, complexity), workflow frameworks
- ❑ Data Access and Interoperability Task Force (DAITF)

# Communities and Data Centers

Which common services are needed?

What are the basic requirements?



# Community Service Wishes

## **In Progress as Services (Task Forces set up)**

- Safe Data Replication (for Bit-stream Preservation & Access Optimization)
- Dynamic Data Replication into HPC Workspace

## **In Specification/Discussion as Services**

- Aggregated EUDAT Metadata Domain
- Researcher Data Store (Simple Upload, Share and Access)

## **In Progress as Research Issues (WP7)**

- more elaborate policy rules and federation scalability
- generic workflow execution framework  
(automatic annotation, data mining, etc.)

All 5 core communities basically share same service wishes  
slightly different priorities

# Enabling Technologies

- **Building robust and available persistent identifier service (is in place based on Handles)**

- EPIC: millions of objects, DataCite: published collections
- EPIC offers registration/resolution service for all data centers in Europe
- EPIC currently 3 strong data centers with redundancy setup
  - will be extended to 10-12 collaborating strong data centers
- EUDAT: all objects need to be registered, all policy operations will use PIDs

ready  
to go

- **Federated AAI service**

- Shib/SAML based world - still a mess due to fragmentation
- can we rely on harmonized EU wide Identity Federation?
- will individual identity providers offer needed attributes?
- do we need to fall back on own user administration?

not yet  
ready

- **Shared Workspaces**

- obviously for different purposes (storing data, automatic annotations, etc)

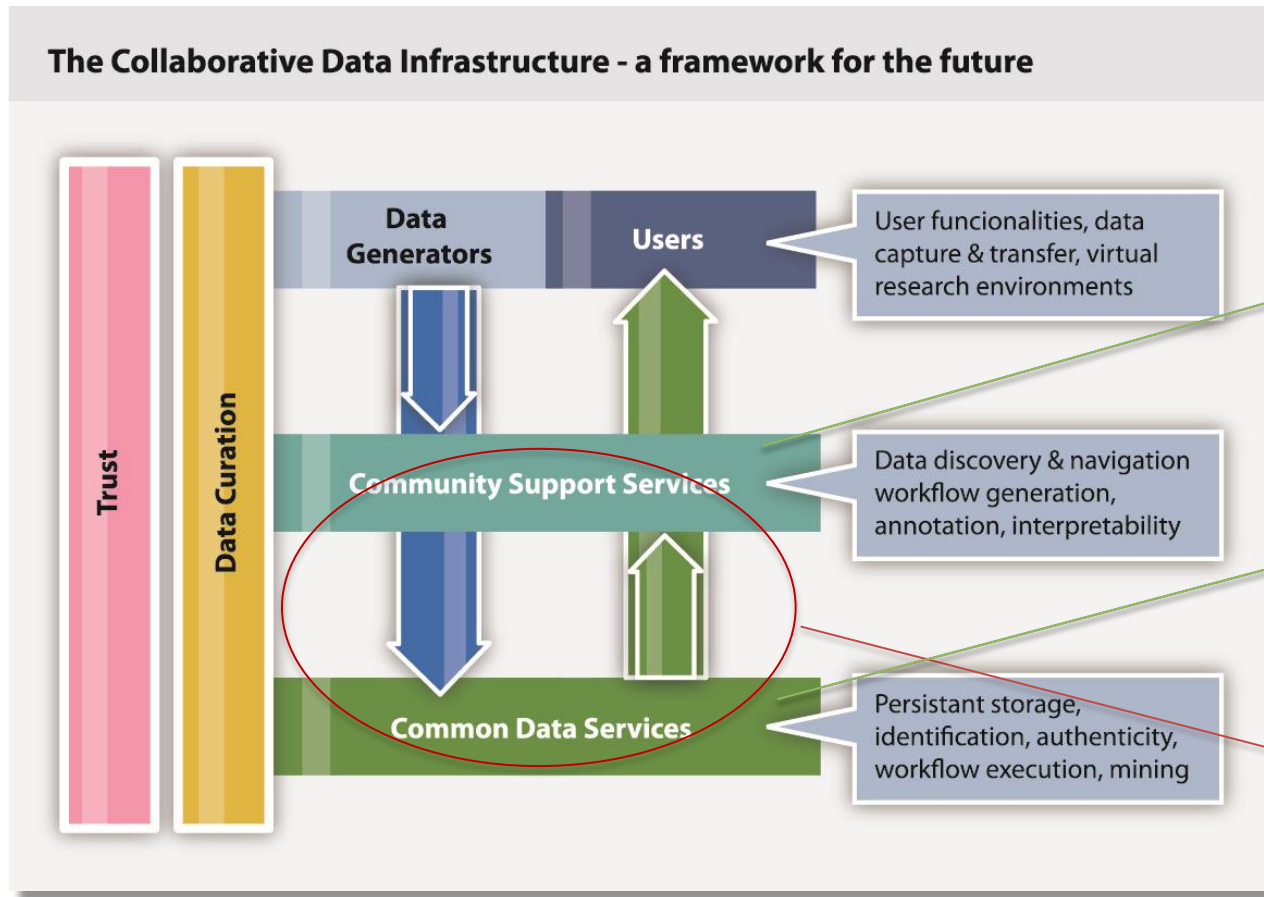
to be  
done

- **Monitoring and accounting**

- all participating servers/services need to show stability, availability

- **Network Services (of course)**

# First - need to understand CDI



CLARIN, LifeWatch, ENES,  
EPOS, VPH, etc.  
5 Core Infrastructures  
~15 second round  
infrastructures

=> 10 EUDAT data centers

indeed some  
heterogeneity at both  
levels

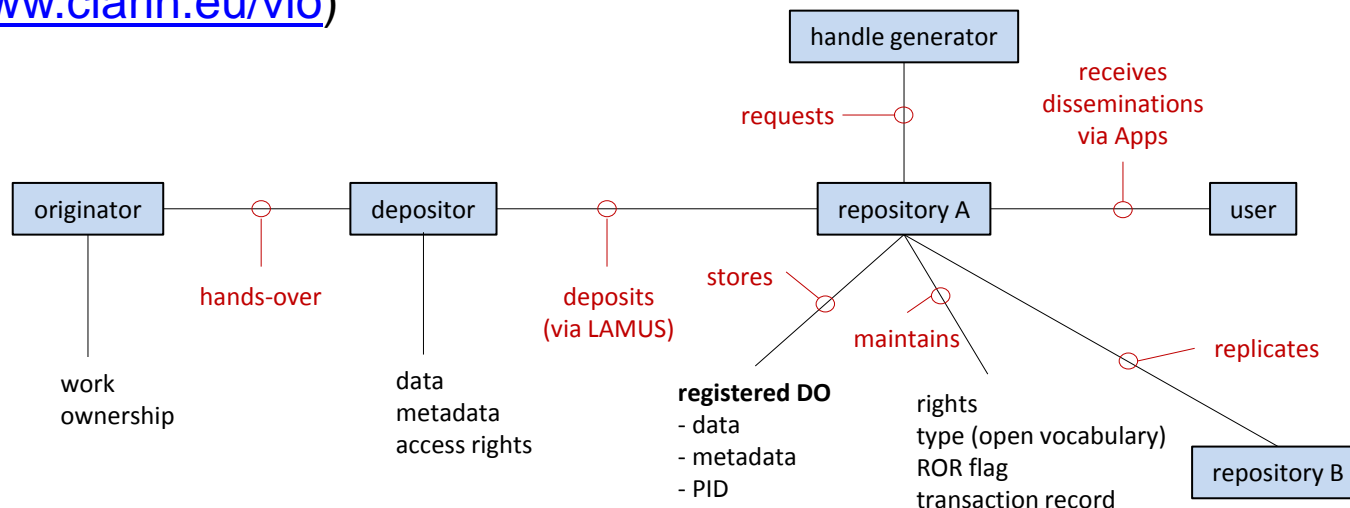
Interviews based on Abstract Data Object Model  
(interviews fostered architecture clarifications)



# Data Landscape Analysis: CLARIN

- **CLARIN (Language Resource and Technology Community)**

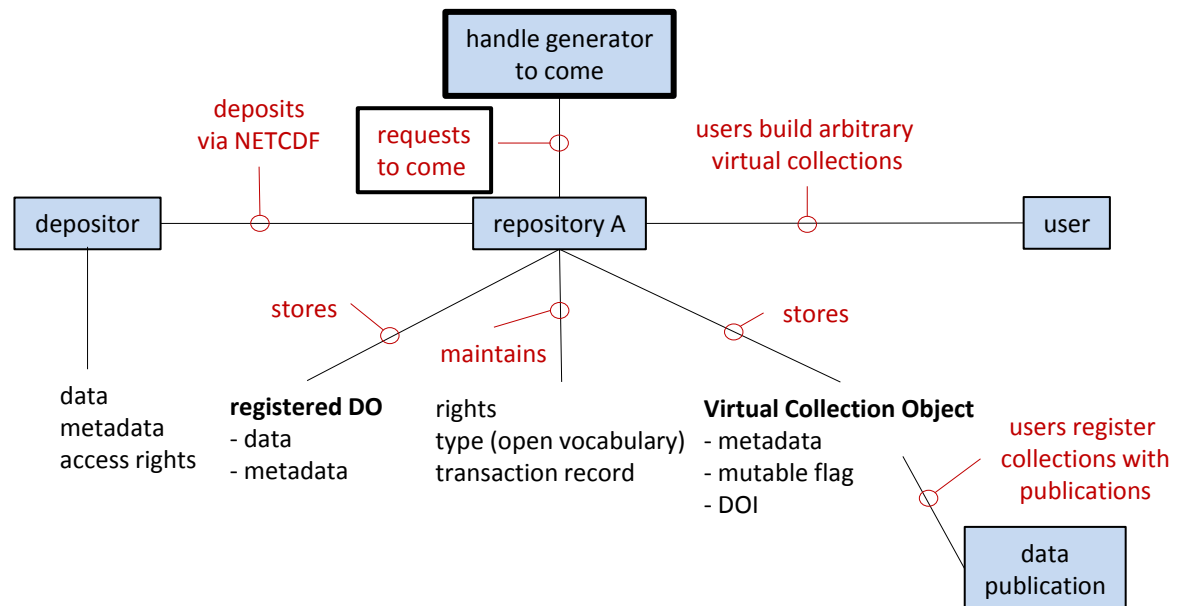
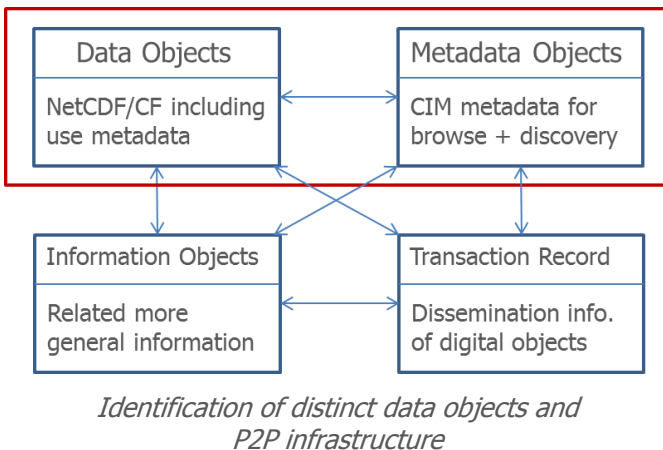
- about 200 centers in Europe with about 30 „community center“ candidates
- have 4 types of centers (DataONE: tiers) from strong to weak requirements
- requirements: rep. system, PIDs, CMDI based metadata, AAI
- almost all busy with re-structuring - only few fulfill strong requirements
- components/profiles and concepts registered (ISOcat, SCHEMcat)
- Virtual Language Observatory: harvesting, mapping, indexing  
([www.clarin.eu/vlo](http://www.clarin.eu/vlo))



# Data Landscape Analysis: ENES

## • ENES (Climate Modeling Research)

- about 20 centers in Europe -
- have CIM data model - but this is still in a prototype state, not deployed broadly
- but CDI as operating at German Climate Center is taken as basis
- CIM has kind of „canonical“ design using DOIs and EPIC Handles
- Metadata based on ISO 11179 etc.; OAI-PMH in place

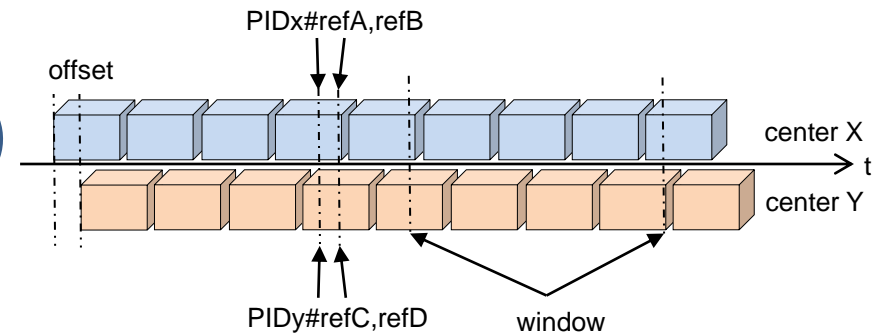
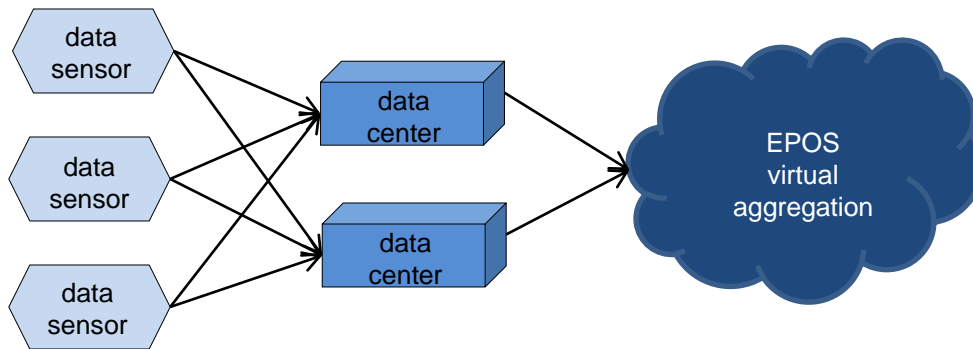




# Data Landscape Analysis: EPOS

- **EPOS (Seismologists, Vulcanologists, etc.)**

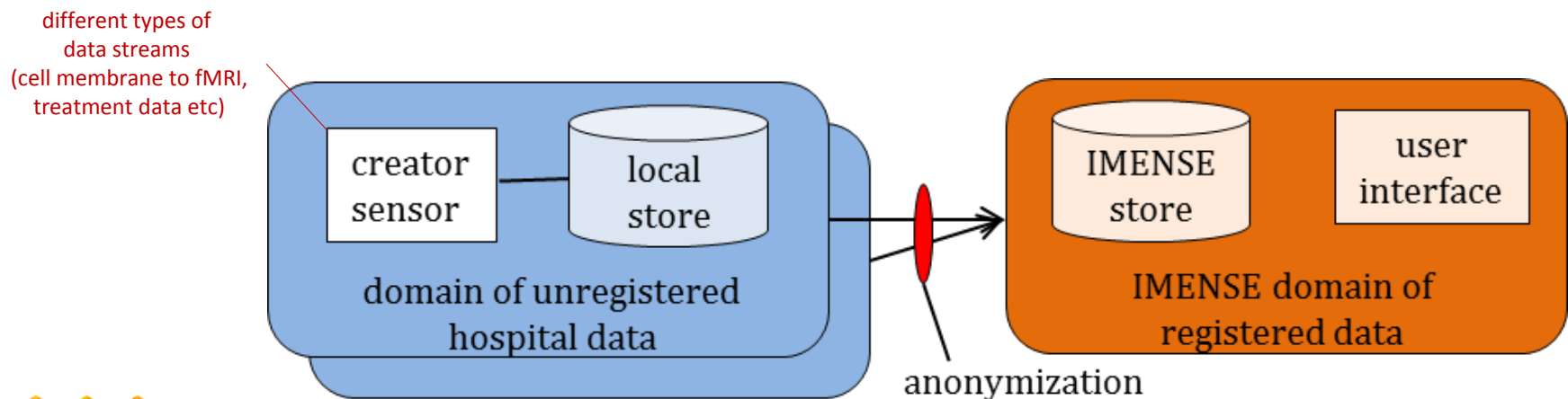
- lots of distributed data sensors producing continuous package streams
- due to various reasons data streams include gaps to be filled over time
- data windows of interest (Wol) are defined „vulcano eruption X“
- aggregations of such data are of relevance (large scale statistics etc)
- work currently on a description of metadata schema for Wols
- work on a scheme of how to refer to packages and offsets (Handles, fragments)
- one center is now implementing reference architecture
- need to synchronize with US and other colleagues



# Data Landscape Analysis: VPH

- **VPH (Virtual Physiology of Humans)**

- currently pilot project with about 5 hospitals in different countries
- one centralized data center - in next phase distributed system
- focus was on metadata aggregation
- IMENSE stores all textual data and Metadata in a DBMS and gives access
- data aggregation is planned together with a large data center in EUDAT
- metadata not yet standardized & formalized (DICOM, JPEG headers, etc.)
- nothing done with PIDs, AAI and OAI-PMH yet





# Data Landscape Analysis: LifeWatch

- **Biodiversity (much based on GBIF)**

- yet no chance of qualified interaction due to time restrictions
- different contributors and actors
- very heterogeneous domain
- first requirements & implementations without LifeWatch
- need to be flexible enough anyhow

# Data Landscape Analysis: 2nd Round

- second round of interviews to come in February/March
- User Forum (March) to meet even other initiatives and start interactions

Environmental Science	ENES, EPOS, Lifewatch, EMSO, IAGOS-ERI, ICOS, Euro-Argo, ...
Social Sciences and Humanities	CLARIN, CESSDA, DARIAH, ...
Biological and Medical Science	VPH, ELIXIR, BBRMI, ECRIN, DiXA, ...
Physical Sciences and Engineering	WLCG, ISIS, DESY, PanData, ...
Material Science	ESS, ...



# Data Landscape Analysis: Summary

- **panta rei - all is moving**

- data infrastructures are shooting on a moving target
  - from core communities only 2 have a ready made architecture
- process of discussion is rather fruitful
  - forces explicitness and fosters harmonization
  - discussions and moderation roles are highly appreciated
- data volumes ready to be contributed range from Exabytes to Terabytes

# Back to community Service Wishes

## **In Progress as Services (Task Forces)**

- Safe Data Replication (for Bit-stream Preservation & Access Optimisation)
- Dynamic Data Replication into HPC Workspace

## **In Specification/Discussion as Services**

- Aggregated EUDAT Metadata Domain
- Researcher Data Store (Simple Upload, Share and Access)

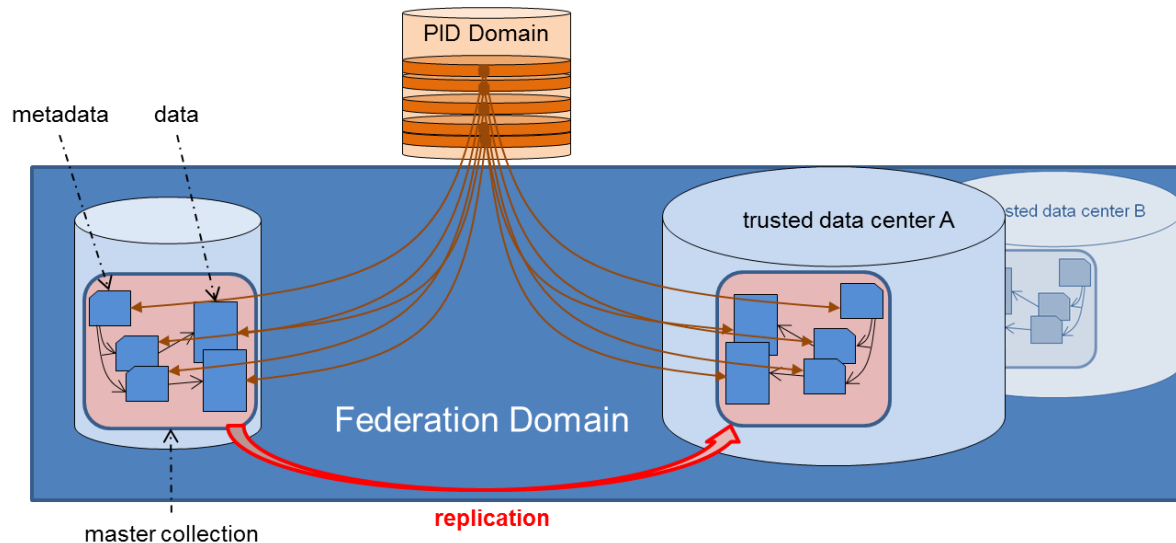
## **In Progress as Research Issues (WP7)**

- more elaborate policy rules and federation scalability
- generic workflow execution framework  
(automatic annotation, data mining, etc.)



# SAFE Data Replication

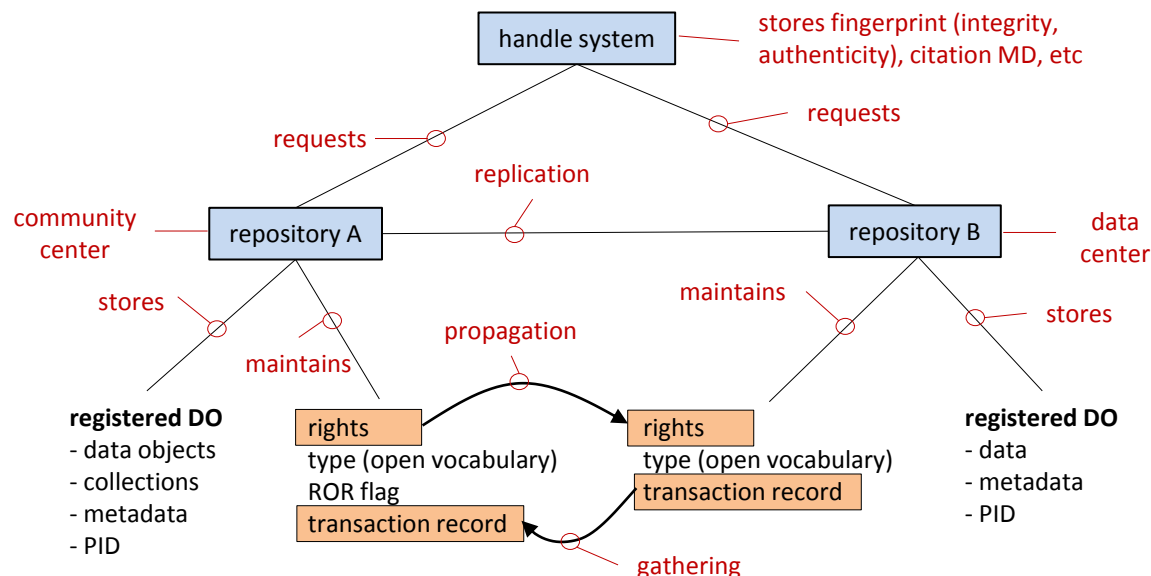
- safe replication between 1 community center and N data centers
- flexibility, scalability and management require policy rule based approach
- 3 islands (community + data center) in parallel & close interaction



- basic technologies: AAI, iRODS, Handles, community MD & OAI-PMH, center registry
- in June merging of 3 islands to one flexible replication domain
- REPLIX experience is basis

# REPLIX

- safe replication between CLARIN center and RZG data center
- purpose: preservation, computation (AV Recognition) and access optimization
- total amount: 80 Terabytes
- requires policy rule based approach due to quality assessment (Data Seal)
- iRODS, Handles, CMDI Metadata
- deployment of Archive/Access software stack as well

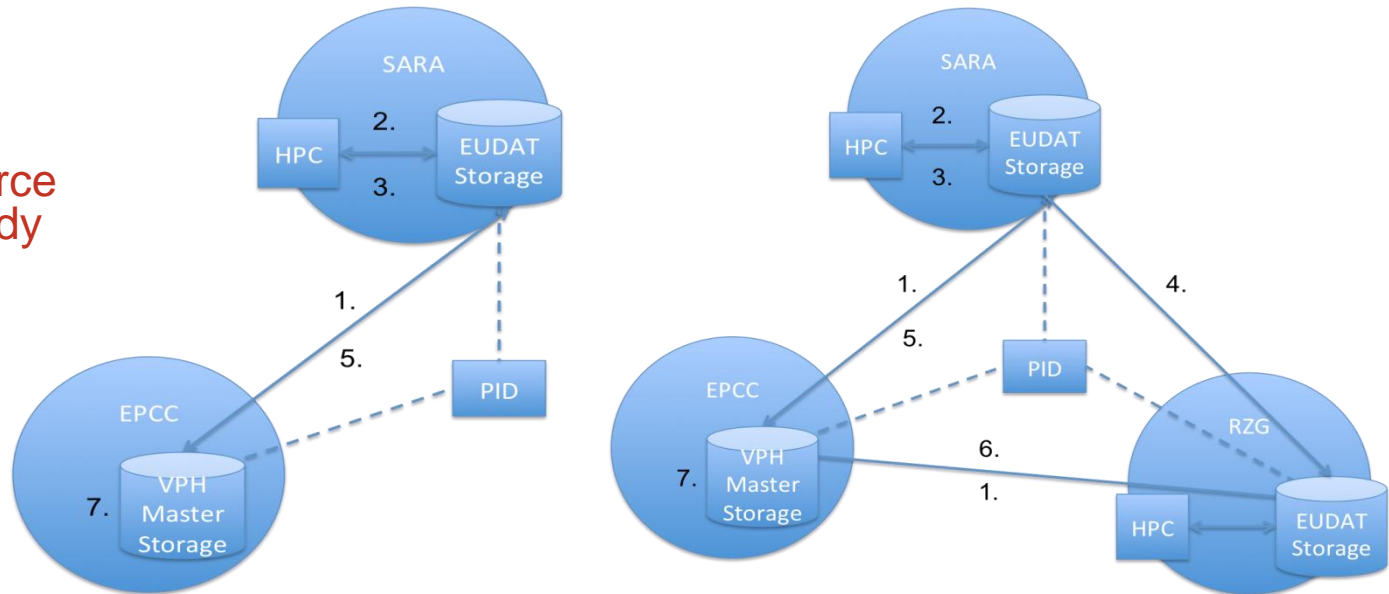


replication at logical collection level basis for demos at ASIST and ICRI conferences both in March (MPI - RENCI)

# Staging to HPC Pipes

- intention is to make use of HPC machines for computations on stored data
- different configurations possible:
  - computations on a single HPC node where data already is
  - computations on multiple nodes - use of PRACE fast distributed file system

Expert Task Force  
built, to be ready  
in summer



- principles:
  - user issues a compute command
  - script pushes data into the HPC workspace, results go into workspace
  - input data is discarded after job end, user needs to store the results

# Aggregated Metadata Domain

- not yet fully specified
- question: for what ???
  - probably loss of specific information - thus interdisciplinary research
  - should show what is stored in the EUDAT data centers
  - one stop shop for virtual collection building
  - making PR for collections (ANDS model)
- general index with some faceted browsing machine probably not sufficient
  - element semantics probably too different
- therefore currently analysis of semantics and simple mapping schemes
- enabling technologies:
  - OAI-PMH, refs via PIDs, SOLR/Lucene for indexing/browsing
  - when and how semantic expansion
  - do we need higher performance technology?
- decision about criteria in February
- technology watch in March

# Researchers Simple Store

- not yet fully specified
- question: for what ???
  - researchers need/want Simple Store for all their „secondary“ data
  - trust is an important issue - owner/copyright must be (with) the researcher
  - data should be part of the EUDAT data domain (thus Metadata, PIDs)
  - ingest via community control to prevent misuse
- Simple Store must have simple access component (like YouTube) and perhaps easy ‚promotion‘ of data into community center collections
- enabling technologies:
  - AAI, PIDs, MD Indexing
- decision about criteria in February
- technology watch in April (what about Mercury etc.)

# EUDAT CDI Summary

- understand data organizations as bottom-up exercise
- determine „common“ functions needed
- determine essential independent components with chance of wide acceptance
  - PID system, center registry, metadata landscape
- define agreed APIs for different components
- rely on policy-rule based approach
- currently implementation of procedures for 3 islands
- probably need to extract common characteristics to scale up
- are looking for close collaboration with others (US, etc.)



# What about DAITF?

- Do we need a **Data Interoperability and Access Task Force**?
- We have already:
  - IETF, W3C, OASIS, OAI, CODATA, GRDI, e-IRG,
  - ISO, ISO/IEC JTC1, ITU, MOIMS (RAC), DSA, etc.
- Many promising initiatives world-wide dealing with the same questions
  - data is global, communities are acting globally
  - much overlap in intentions - however slight differences
- Our conclusion: we need a forum (whatever name we give it) where
  - “data practitioners” can meet regularly - no PR, no politics
  - we can exchange approaches & technologies, discuss harmonization, standards, IT principles, etc.
  - we can train young “data scientists”
- Are we already strong enough to go outside?

# What about DAITF?

- who: data architects, data practitioners, information experts, ?
- what first:
  - need to define the scope
    - there is so much community specific “pre-registration” activity
  - need to meet with a prepared agenda
  - need to have a start-up Steering Group and perhaps first WGs
- agenda:
  - first discussion at DAITF Preparation Workshop at ICRI in Copenhagen - March 20/21.
  - EUDAT/OpenAIRE received money to host 2 Workshops 12/13
  - submitted an application to EC with DAITF continuation
- for EC Data Infrastructures will be a top priority in Horizon 2020
  - EC is going to continue funding DAITF